REED ON THE COEFFICIENT OF CORRELATION.

Since the development of the modern higher statistical methods has been largely in the service of biology, it is only natural that those who adopt the biometrician's formulae without a first hand acquaintance with the extensive biological literature should often labor under the disadvantage of using methods of calculation which are not the best suited to the practical needs of their work. If the methods adopted really give accurate constants, the use of more laborious routine processes is merely the misfortune of the individual worker. When, however, the beginner attempts, as is surprisingly often the case, to elucidate the subject for others, real harm may result, particularly when cumbersome and laborious methods are proposed in cases in which simple and direct formulae are already available.

A case in point is afforded by a recent paper by Ree

(1) The arithmetical routine indicated by Reed for the determination of the correlation coefficient by the summation of actual deviations, squares of deviations, and products of the deviations of the two variables from their respective means is wholly unsuited for practical work. All of the labor involved in the calculation of the five columns of his Table I is quite unnecessary. The elimination of this superfluous routine not only saves time but minimizes the danger of error which is particularly great when signs must be constantly regarded. The moments may be taken in terms of units about any point whatever. In these days of splendid calculating machines permitting simultaneous multiplication and summation, deviation methods involving positive and negative signs are inferior to one in which the first and second powers and the products of the. two variables are summed, either directly from unclassified data or from the entries in a conventional correlation table. The formulae† required are:

$$\bar{x}=\Sigma(x)/N \qquad \bar{y}=\Sigma(y)/N$$

$$\sigma_x=\sqrt{\Sigma(x^2)/N-[\Sigma(x)/N]^2} \qquad \sigma_y=\sqrt{\Sigma(y^2)/N-[\Sigma(y)/N]^2}$$

$$r_{xy}=\frac{\Sigma(xy)/N-\bar{x}\bar{y}}{\sigma_x\,\sigma_y}$$

where x and y are the actual measures of the two variables, not their deviation from their mean value, the bars denote means and the sigmas standard deviations for the whole population of N individuals.

By the direct summation method Reed's Table I gives:

For July Rainfall, x, in inches:

$$\Sigma(x) = 243.9 \qquad \Sigma(x)/N = 4.065$$

$$\Sigma(x^2)=1103.87 \qquad \Sigma(x^2)/N=18.3978$$

$$\sigma_x= \sqrt{\Sigma(x^2)/N-[\Sigma(x)/N]^2}=1.3687$$

*Reed, W. G. The Coefficient of Correlation. Quart. Pub. Amer. Stat. Asso., Vol. 15, pp. 670–684, 1917.

†Harris, J. Arthur. The Arithmetic of the Product Moment Method of Calculating the Coefficient of Correlation, Amer. Nat., Vol. 44, pp. 693–699, 1910.

For yield of corn in bushels, y,

$$\Sigma(y) = 2067.1 \qquad \Sigma(y)/N = 34.45$$
$$\Sigma(y^2) = 72446.69 \qquad \Sigma(y^2)/N = 1207.445$$
$$\sigma_y = 4.531$$

For rainfall and yield

$$\Sigma(xy) = 8612.99, \Sigma(xy)/N = 143.5498$$
$$r_{xy} = \frac{\Sigma(xy)/N - \bar{x}\bar{y}}{\sigma_x \sigma_y} = .565 \pm .059.$$

Reed gives $r = .526 \pm .063$, a value which is slightly erroneous because of the approximations necessary in the cumbersome method of calculation employed.

Full details concerning the application of this method to correlation tables, for both integral and graduated variates, are given in the paper cited.

(2) Reed lays entirely too great emphasis upon the necessity for a preliminary testing of linearity of regression and his discussion of the methods is very misleading. He says: "The correlations should never be attempted without first investigating the relationship far enough to see if it follows a straight line."

Now while it is true that the correlation coefficient is not strictly valid except in cases of linear regression it always gives at least a minimum measure of the relationship between two variables, and in the vast majority of cases this measure will be the one finally adopted after linearity has been tested.

Furthermore, the determination of the coefficient of correlation *is the first step in the critical testing of linearity.* It is quite unnecessary to plot the two variables in units of their standard deviation or to use the method of least squares in adjusting the line as suggested by Reed. When the correlation coefficient is known the straight lines are simply:

$$x = (\bar{x} - r_{xy}\frac{\sigma x}{\sigma y}\bar{y}) + r_{xy}\frac{\sigma x}{\sigma y}y,$$

$$y = (\bar{y} - r_{xy}\frac{\sigma y}{\sigma x}\bar{x}) + r_{xy}\frac{\sigma y}{\sigma x}x.$$

These can be calculated without even tabulating the raw data. Thus the equation from Reed's Table I is

$$y = 26.849 + 1.870\,x,$$

where y = yield in bushels and x = July rainfall in inches. This equation shows at once that the line in Reed's Diagram 2 on page 673 is incorrect. It should indicate an average yield of 30.59 bushels for 2 inches and 41.8 for 8 inches of July rainfall.

When the correlation coefficient and the regression equation have been determined, linearity may be tested by inspection of the distribution of the actual frequencies represented on a scatter diagram like Figure 2 or 3 of

Reed's paper, by a comparison of the empirical means of arrays with the theoretical means calculated from the regression equations, or by the use of Blakeman's criterion.*

Thus the calculation of the correlation coefficient, which by a practical method is a very easy task, should always be the first step in the analysis of the data. The nature of the regression line can then be determined in a convenient and really scientific manner.

(3) The bibliography given by Reed as "containing complete statements of the later development of work on the theory of correlation" is far from complete. A partial list of the omissions may be found elsewhere.†

This note is written with no desire to criticize the work of an individual writer, but merely in the hope of saving some beginner, who might follow Reed's suggestions, excessive and unnecessary labor in the calculation of the correlation coefficient, or misconceptions concerning the value or applicability of this most important statistical constant.

Cold Spring Harbor, N. Y. J. ARTHUR HARRIS.

COMMENT ON PROFESSOR IRVING FISHER'S ARTICLE ON THE "RATIO CHART."

The following remarks upon Professor Fisher's article in the *Quarterly Publications* for June are made without controversial intent and published rather as supplementary to than as critical of the statements made in the article in question. Where there is entire agreement, however, there is naturally little occasion for comment.

In the matter of terminology I do disagree with Professor Fisher, but that is hardly worth debating. The choice of the name "ratio chart," and the suggested opposition between "ratio" and "difference," is open to objection. Are not the curves in question currently and correctly known as logarithmic? Is it worth while to avoid this term simply because it, or the conception involved, is not familiar to the man in the street? All statisticians are acquainted with such curves, though many, it is true, feel them to be alien, just as some of them feel all graphic devices to be alien to their work. Should not the re-naming of a thing to make it popular be left to the commercial classes as a suitable mercantile, rather than a scientific or properly pedagogic, device? But, if the change means an improvement in our terminology, as regards either accuracy or adequacy, the case is different.

According to the usage with which I am familiar, statistical curves, and the diagrams constituted by them, are either arithmetic or logarithmic. Numbers and quantities are absolute or relative. Differences, also, may be absolute or relative. A relative difference is a species of ratio, and the

*Blakeman, J. On Tests for Linearity of Regression in Frequency Distributions. Biometrika, Vol. 4, pp. 332-350. 1905.

†Harris, J. Arthur. An Outline of Current Progress in the Theory of Correlation and Contingency. Amer. Nat., Vol. 50, pp. 53-64, 1916.